⑲ 日本国特許庁（JP）　　⑪ 特許出願公開

⑫ 公開特許公報（A）　　昭59—139099

⑤Int. Cl.³	識別記号	庁内整理番号	㊸公開　昭和59年(1984)8月9日
G 10 L 1/00		R 7350—5D	

発明の数　1
審査請求　未請求

（全 4 頁）

⑤④音声区間検出装置

㉑特　　願　昭58—13997
㉒出　　願　昭58(1983)1月31日
㉒発 明 者　坂田富生
青梅市末広町2丁目9番地の1
東京芝浦電気株式会社青梅工場内
㉑出 願 人　株式会社東芝
川崎市幸区堀川町72番地
㉔代 理 人　弁理士　鈴江武彦　　外2名

明　細　書

1. 発明の名称

音声区間検出装置

2. 特許請求の範囲

入力音声信号から音声パラメータ時系列を抽出する音声パラメータ抽出手段と、この音声パラメータ抽出手段から出力される上記音声パラメータ時系列を一時的に格納するバッファメモリと、上記音声パラメータ時系列に基づいて背景雑音の音声パラメータ値の平均値を算出する雑音レベル計算手段と、この雑音レベル計算手段から出力される上記背景雑音の音声パラメータ値の平均値に基づいて決定されるバイアス値を含む音声区間検出用閾値を算出する閾値計算手段と、上記音声区間検出用閾値に基づいて上記バッファメモリに格納された音声パラメータ時系列から音声区間の始端および終端の両者をそれぞれ検出する音声区間検出手段とを具備したことを特徴とする音声区間検出装置。

3. 発明の詳細な説明

〔発明の技術分野〕

この発明は、音声認識システムに使用される音声区間検出装置に関する。

〔発明の技術的背景とその問題点〕

音声認識システムにおいては、その前処理として音声区間の検出を正確に行なう必要がある。通常、音声区間検出において、信号対雑音比（S/N比）が良好で（例えばエネルギーS/N比にして30dB以上の音声波を扱う場合）、しかも背景雑音レベルがあまり変動しないような環境下では比較的容易に検出を行なうことができる。具体的な検出方式としては、音声波を広帯域マイクロホンを介して入力し、その入力音声信号の短時間エネルギーおよび電交差数を求め、これらが所定の固定閾値を所定期間連続して越えるか否かを調べるなどの方式がある。

このような固定閾値方式では、背景雑音レベルが時間的にある程度変動する場合には下記のような問題が生ずる。即ち、まず固定閾値が低

く設定されると、背景雑音レベルが少し高くなっただけで、閾値を越えてしまい雑音を音声区間の一部として取込むという不都合がある。逆に固定閾値が高く設定されていると、音声区間中のレベルの低い部分を取りこぼすという不都合がある。このような点を解決するためには、背景雑音レベルに応じた閾値を決定する方式がある。即ち、まず音声が発声される前（後）の無音区間と見なされる区間での入力音声信号の短時間エネルギーおよび零交差数の平均値を求める。そして、この平均値に所定の固定バイアス値を加えたものを閾値として用いることが行なわれる。

しかしながら、上記のような方式においても、背景雑音レベルの変動が大きい場合には、固定バイアス値による閾値では正確な音声区間検出は困難である。これは、仮にバイアス値を低く設定すると、短時間エネルギーおよび零交差数が閾値を越える雑音区間が頻出することになる。これにより、雑音区間が音声区間の一部として取込まれたり、または雑音区間のみが音声区間として検出されるという重大な誤動作が生ずる。逆に、バイアス値を高く設定すると、音声区間の一部または全部が欠落するという誤動作が生ずる欠点があった。

〔発明の目的〕

この発明は上記の事情に鑑みてなされたもので、その目的は、背景雑音レベルの変動が大きい場合でも、適切なバイアス値を加えた閾値を設定することにより、正確な音声区間検出を行なうことができる音声区間検出装置を提供することにある。

〔発明の概要〕

この発明は、入力音声信号に基づく音声パラメータ時系列より音声区間の始端および終端を検出する音声区間検出手段を用いる。この場合、音声パラメータ時系列により、音声信号の入力直後の数フレームにおける無音区間の音声パラメータの平均値を雑音レベル計算手段で求める。さらに、無音区間の音声パラメータの平均値に基づいて決定されるバイアス値を含む閾値が、閾値計算手段により算出される。この閾値に基づいて、音声区間検出手段は音声区間の始端および終端を検出するものである。

〔発明の実施例〕

以下図面を参照してこの発明の一実施例について説明する。第1図はこの発明に係る音声区間装置の構成を示すブロック図で、1は音声信号Sの入力端子である。音声信号Sは、入力端子1から音声パラメータ抽出部2に与えられる。音声パラメータ抽出部2は、音声信号Sから短時間エネルギー等の音声パラメータ時系列を抽出する。バッファメモリ3は、音声パラメータ抽出部2から出力する音声パラメータ時系列を一時格納する。4は音声区間検出部で、閾値計算部6から出力する閾値E_{TH}に基づいて、バッファメモリ3からの音声パラメータ時系列における音声区間の始端および終端を検出する。閾値計算部6は、雑音レベル計算部5から出力する背景雑音の音声パラメータ値の平均値に応じて決定されるバイアス値を含む音声区間検出用の閾値E_{TH}を算出して出力する。雑音レベル計算部5は、バッファメモリ3の音声パラメータ時系列より、音声信号Sの入力開始直後の数フレームにおける無音区間の音声パラメータの平均値（背景雑音の音声パラメータ値の平均値）を算出して出力する。7は出力端子で、音声区間検出部4で得られた音声区間の始端および終端の情報を出力する。

このような構成において、その動作を説明する。音声信号Sは、通常広帯域マイクロホンまたは電話回線等を介して音声パラメータ抽出部2に与えられる。音声パラメータ抽出部2は、音声信号Sの1フレーム間のrms値、即ち短時間エネルギーEを各フレーム毎に計算し出力する。ここで、フレーム幅及びフレーム周期は10msec程度とする。このようにして、音声パラメータ抽出部2は、第2図に示すような音声パラメータ時系列をバッファメモリ3に出力する。雑音レベル計算部5は、バッファメモリ3

から第1フレーム（即ち音声信号Sの入力開始時点）から第Mフレームまでの音声パラメータ値を読み出す。そして、Mフレーム（例えば80～100msec）の音声パラメータの平均値E_Iを求める。この平均値E_Iは、背景雑音の音声パラメータ値の平均値としてみなされる。これは、一般に音声認識装置ではディスプレイおよび信号音等により発声者に発声タイミングを知らせ、同時に信号を取込み始める。しかしながら、通常、発声者は発声促進信号と同時に発声するということは殆んどなく、発声促進信号が出力された後、少し遅れて発生する。したがって、音声信号Sの入力開始後100msec程度は、無音区間であると考えられ、E_Iは背景雑音の音声パラメータ値の平均値とみなされる。

上記のようにして、雑音レベル計算部5から出力される背景雑音の音声パラメータの平均値E_Iは、閾値計算部6に与えられる。閾値計算部6では、平均値E_Iに基づいて第3図に示すE_I－バイアス値αの関係よりバイアス値αを求め、「$E_I+\alpha$」を音声区間検出用閾値E_{TH}として出力することになる。ここで、第3図は下記のような式(1)～(3)により求められる。

$$E_I < E_1 \text{ で } \alpha=\alpha_1 \quad \cdots\cdots(1)$$

$$E_1 \leqq E_I < E_2 \text{ で } \alpha=\frac{\alpha_2-\alpha_1}{E_2-E_1}(E_I-E_1)+\alpha_1 \quad \cdots(2)$$

$$E_I \geqq E_2 \text{ で } \alpha=\alpha_2 \quad \cdots\cdots(3)$$

これは、一般に音声信号入力部の増幅器または電話回線系の利得がn倍になったとすると、音声と共に背景雑音の短時間エネルギーの平均値及び分散はn倍になる。音声区間検出用閾値E_{TH}を、背景雑音の短時間エネルギーの平均値E_Iにバイアス値αを加えた形で与えることにすると、第1項E_Iにより平均値の変動は吸収されるが、分散の変動は吸収されない。そこで、分散の変動を、第2項のバイアス値αを適切に設定して吸収することになる。そのため、上記式(2)に示すようにバイアス値αをE_Iの値に応じて線形に変化させればよい。これが$E_1 \leqq E_I < E_2$の区間である。但し、E_Iが極端に小さい（大きい）場合には、上記式(2)で計算されるバイアス値が小さく（大きく）なり過ぎることを避けるため、式(2)を適用するE_Iの範囲に制限を付け、$E_I < E_1$，$E_I \geqq E_2$の範囲ではそれぞれ固定値α_1，α_2とする。ここで、E_1，E_2，α_1，α_2等の値は実験的に設定される。

音声区間検出部4は、閾値計算部6で算出された閾値E_{TH}に基づいて、バッファメモリ3から読出す音声パラメータ時系列における音声区間の始端aおよび終端bをそれぞれ検出する（第2図）。具体的には、まず始端aの検出において、音声信号の入力開始時点から短時間エネルギーEの時系列を辿り、最初に$E > E_{TH}$となる時点$\tilde{a}$を検出する。この$\tilde{a}$により、$E > E_{TH}$なる区間、即ち音声区間が所定フレーム数N_1だけ継続するか否かを調べる。このフレーム数N_1は例えば50～60msecに相当する値である。そして、N_1フレーム継続の条件が満足されたとき、上記時点$\tilde{a}$を始端aとして出力する。また、$\tilde{a}$以降、$E > E_{TH}$なる区間がN_1フレーム継続しないときにはこれを雑音によるものとみなして、改めて$\tilde{a}$の検出を行なう。

一方、終端bの検出において、始端aより音声パラメータ時系列を辿り、最初に$E \leqq E_{TH}$となる時点$\tilde{b}$を検出する。この$\tilde{b}$より、$E \leqq E_{TH}$なる区間が所定フレーム数N_2だけ継続するか否かを調べる。このフレーム数N_2は例えば250～300msecに相当する値である。そして、N_2フレーム継続の条件が満足されたとき、上記時点$\tilde{b}$を終端bとして出力する。なお、$\tilde{b}$以降、N_2フレーム内に$E > E_{TH}$なる区間が出現したとき、その区間が所定フレーム数N_3に達しないならば、これを雑音によるものとみなし、この区間のフレーム数を無音区間のフレーム数に加える。ここで、フレーム数N_3は例えば40～50msecに相当する値である。また、$E > E_{TH}$なる区間がN_3以上継続した場合には、音声区間の別の部分が出現したものとみなして、改めて$\tilde{b}$の検出を行なう。このようにして、音声区間検出部4により音声パラメータ時系列から音声区間

の始端aおよび終端bのそれぞれが検出されて、出力端子7に出力される。

〔発明の効果〕

以上詳述したようにこの発明によれば、背景雑音レベルが高く、しかも時間的に大きく変動する場合でも、背景雑音の音声パラメータの平均値に基づいて求められる適切なバイアス値を含む閾値を設定し、この閾値に基づいて音声信号の音声区間における始端および終端を確実に検出できる。したがって、音声信号の音声区間検出を正確に行なうことができ、結果的に音声認識システムの精度を向上できるものである。

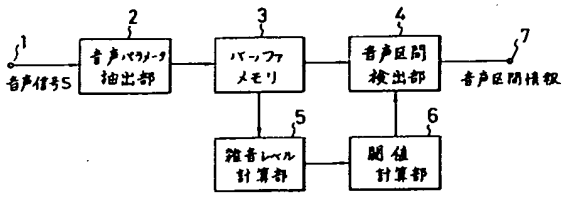
4. 図面の簡単な説明

第1図はこの発明の一実施例に係る音声区間検出装置の構成を示すブロック図、第2図は第1図の装置の動作を説明するための音声パラメータ時系列の波形を示す図、第3図は第1図の装置の動作を説明するための平均値E_I－バイアス値αの関係を示す図である。

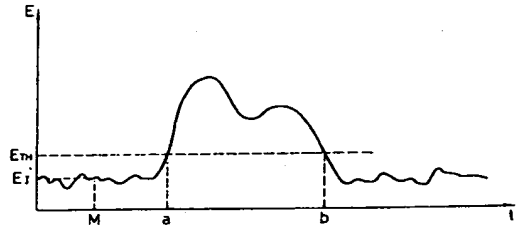
2…音声パラメータ抽出部、3…バッファメモリ、4…音声区間検出部、5…雑音レベル計算部、6…閾値計算部。

出願人代理人　弁理士　鈴　江　武　彦

第1図



第2図



第3図